The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

カテゴリの親子関係の種類に基づくWikipediaカテゴリの再整理

Reorganizing Hierarchical Category Structure of Wikipedia Based on the Parent-Child Relationship Classification

中川嵩教 *1　小板橋佳晃 *1　吉岡真治 *1*2
Takanori Nakagawa　Yoshiaki Koitabashi　Masaharu Yoshioka

*1北海道大学 Hokkaido University　*2理研 AIP RIKEN AIP

Wikipedia is a largest online encyclopedia that covers varieties of topics using structured documents (e.g., infobox for describing metadata, and classification using Wikipedia category). There are several efforts to extract structured knowledge, such as DBpedia, YAGO2, and Japanese Wikipedia ontology. However, there is no research that try to analyze whole Wikipedia category hierarchical structure. To tackle this problem, we have been working on the project to reorganize Wikipedia category hierarchical structure for knowledge engineers and proposed classification of parent-child relationships (e.g., class-subclass and class-instance) in the Wikipedia category hierarchy. In this paper, we discuss effectiveness of this classification results for Japanese Wikipedia category hierarchy by analyzing the characteristics of reorganized category hierarchies that are constructed by selecting (or removing) particular types of relationship.

1. はじめに

Wikipedia[*1] は世界最大のインターネット百科事典であり、その特徴として、ページ、インフォボックス、カテゴリといった構造化がなされている。このように構造化された情報を用いて、DBpedia[Bizer 09] では、ページの情報に関するメタデータの抽出や、カテゴリ階層をオントロジーの構築に利用する YAGO2[Hoffart 13] や、日本語 Wikipedia オントロジーの研究 [玉川 10] などが行われている。しかし、確かに Wikipedia カテゴリには、概念階層としてみなせる階層が存在するものの、単純な包含関係ではない階層関係が存在し、単純に Wikipedia カテゴリの階層全てを概念階層として扱うと不都合が生じるため、既存の研究 [Hoffart 13, 玉川 10] では、アドホックに、その一部のみを利用していた。この問題に対し、Wikipedia カテゴリを分析する研究 [Yoshioka 14, 藤原 12] から、カテゴリには概念分類を表すものだけでなく、人名や組織名といったインスタンスを表すものやその組み合わせが存在しているために、知識工学的観点から利用するためには、そのカテゴリ間の関係について検討することが必要であることが提案された [中川 18]。また、この考え方に基づき、カテゴリの種類、カテゴリの親子関係の種類を整理したデータの構築を行った [中川 19]。本稿では、この構築したデータに基づき、特定の親子関係のみを利用して (あるいは、排除して) できあがるカテゴリ階層構造の性質を議論することで、これまでに提案してきた Wikipedia カテゴリに関する分類の枠組の有用性について議論する。

2. Wikipedia カテゴリ

2.1 Wikipedia カテゴリの階層構造

Wikipedia において、カテゴリとは、膨大な記事群を様々な観点から分類するための索引であり、各記事には、それぞれに一つ以上のカテゴリが付与される。また、このカテゴリは、さらに詳細なカテゴリと関連付けることにより、カテゴリは階層的な構造となっている。このカテゴリ階層については、基本的には、下位カテゴリに属する記事は、上位カテゴリの性質も含むという包含関係が成立することが期待される。よってカテゴリ階層は、知識工学で用いられる概念階層と似た性質を持つことが期待されている。しかし、このカテゴリ階層の構造は、Wikipedia に登録される記事の増加に伴い、既存の階層の中に便宜上のカテゴリが作られ、それにより、必ずしも、包含関係が成り立たない形でカテゴリ階層が作られるようになっている。次小節では、包含関係に注目した際に注意すべきであるカテゴリの種類について述べていく。

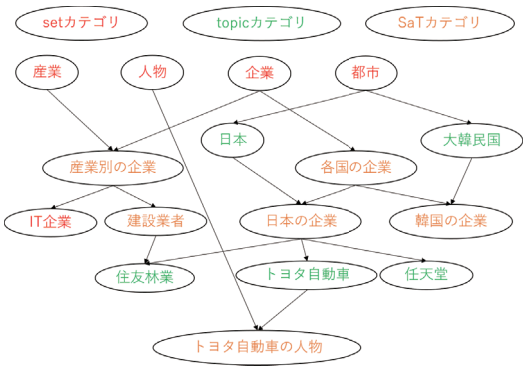

図 1: 分割のためのカテゴリによるカテゴリ分割

2.2 カテゴリの種類

Wikipedia カテゴリは、ページを分類する基準であり、「企業」、「人物」といった概念の種類を表すようなカテゴリだけでなく、「日本」「トヨタ自動車」といった具体的な事象を表すようなカテゴリが存在する。英語版 Wikipedia では、前者を set カテゴリ、後者を topic カテゴリと読んでいる。また、Wikipedia では、一つのカテゴリに大量のページが所属するとそのリストを閲覧することが困難になることから、このようなカテゴリは、様々な基準により、より詳細なカテゴリに分割することが求められている。この時できあがるカテゴリの多く

連絡先: 中川嵩教，北海道大学工学部情報エレクトロニクス学科，札幌市北区北 14 条西 9 丁目,011-706-7161，f-b-hawk78@eis.hokudai.ac.jp

*1 https://www.wikipedia.org/

は、前述の Set と Topic の組合わせであることから、このようなカテゴリは、Set-and-Topic(SaT) カテゴリと呼ばれている。例えば、図 1 の例ではオレンジ色の「日本の企業」「産業別の企業」などが SaT カテゴリに該当する。

3. Wikipedia カテゴリと階層関係の分類

[中川 19] では、Wikipedia カテゴリを概念の種類を表すようなカテゴリを set、具体的な事象を表すようなカテゴリを topic、そして、SaT カテゴリの中で、set を構成要素として持ち、set の性質を持つ制約付き set を CS カテゴリ (ConstraindSet カテゴリ) とし、全てのカテゴリをこの 3 種類に分類した。また、カテゴリの繋がりを包含関係の観点から以下のように分類した。

- 「制約詳細化」: CS → CS である set 部分を共通として topic が詳細化される関係である。例としては「アジアの企業」→「日本の企業」などである。制約付加と同じように set 部分は変わらないので包含関係は満たされる。
- 「クラス-サブクラス」: set 部分を見たときに、同じではないが、概念としては同じとなるような関係である。set → set 関係の「作家」→「著作家」のようなものや、CS → CS 関係の「日本の企業」→「日本の多国籍企業」のような、topic 部分がないまたは共通として、set 部分がクラス-サブクラスとなっている関係が主だが、「SF 作品」→「未来を題材とした作品」や、「日本のクラブに所属するサッカー選手」→「ベガルタ仙台の選手」のように、set 部分だけを見ると逆転が起こっているが全体を見るとクラス-サブクラスとなっているような関係も一定数存在した。概念としては同じということで、この関係においても包含関係は満たされる。
- 「クラス付加」: あるカテゴリに新たな set が付け加わり CS となる関係である。多くは「トヨタ自動車」→「トヨタ自動車の人物」のように topic に set が付け加わる関係が多いが、「アニメ」→「アニメに関する企業」や「歴史の人物」→「歴史の人物を題材とした作品」のように、set 部分の後ろに新たな set が付け加わり、概念が変わってしまう set → CS 関係や CS → CS 関係も存在した。この関係においては上位カテゴリ（topic ならその上位カテゴリ）と下位カテゴリの間では概念が異なるため、包含関係は満たされない。
- 「Instance of」:「格闘技漫画」→「北斗の拳」や「日本の国公立大学」→「北海道大学」といった、下位カテゴリが上位カテゴリの概念の具体例となるような関係である。
- 「topic 包含関係」:「イギリス」→「イングランド」地理的な包含関係があるものが存在する他、「AKB48」→「前田敦子」のようなメンバー関係など、概念の記述はないが、topic 同士で包含関係があるような関係である。

上記に当てはまらないものは例外としてあり、例として、「日本」→「日本関連一覧」のような一覧を表示させるためのカテゴリや、「スポーツ施設」→「スポーツ施設の画像」のような画像を表示させるためのカテゴリなど、Wikipedia 特有と思えるカテゴリ名を含む関係や、歴史や二国間関係等、サブクラスやインスタンスを広く取るものが挙げられていた。

4. Wikipedia カテゴリの再整理

構築した Wikipedia カテゴリならびにカテゴリの親子関係の分類の有用性を検討するために、次の 2 つの観点からの分析を行った。

- 循環構造に関する検討
  Wikipdia カテゴリの階層構造が、親子関係の包含関係 (子カテゴリに属するページは親カテゴリにも属する) を前提としているのであれば、カテゴリ階層に循環構造 (子カテゴリをたどっていくと、親カテゴリに戻る) が存在してはいけないことになるが、現実には、このような構造が Wikipedia には存在する。前節で提案した親子関係の分類には、必ずしも、包含関係が成り立たない関係が含まれていたため、包含関係が成り立つ関係に限定した際に、循環構造の数がどのように変化するかを分析する。
- クラス階層の分析
  Wikipedia のカテゴリは、概念分類が同じであっても、その制約が詳細になっていくことによって、カテゴリの段数が増えていく (「日本の?」→「北海道の?」→「札幌市の?」)。しかし、これらは、本来、一つの概念分類に属するカテゴリを分割していくことでつくられたカテゴリ階層であり、概念分類という意味では、同一に扱うという考え方もある。このような分割のためのカテゴリの影響を除いて、純粋にクラス-サブクラスの概念階層にのみ注目した場合に、できあがる階層について分析する。

4.1 循環構造に関する検討

循環構造は、Wikipedia カテゴリの親子関係を有向グラフとして考えたときの強連結成分として、抽出することができる。今回構築にしようした Wikipedia カテゴリの階層構造に対して、強連結成分の抽出アルゴリズムを適用したところ、58 件の循環構造が見つかった。

このカテゴリ階層には、「ロンドン」→ (クラス付加) →「ロンドンの地方自治」→ (Instance of) →「グレーター・ロンドン」→ (topic 包含関係) →「ロンドン」といった、必ずしも包含関係が保証されないようなクラス付加といった関係を含んでいるものや「東芝」→ (その他) →「東芝グループ」のように、「その他」として分類されている分類不十分なインスタンス間の関係を含むものが存在した。

これらの問題に対し、本研究で提案している分類のうち、包含関係が成り立つと考えている「制約詳細化」、「クラス-サブクラス」に限定して再構築した Wikipedia カテゴリ階層に限定することで、多くの循環構造については、その循環がなくなり、6 件の構造のみが残った。

この 6 件に含まれているものは、、「系譜学」→ (クラス-サブクラス) →「系図」→ (クラス-サブクラス) →「系譜学」のように関係を付け直した方が良いと思われる関係と「中国の鉄道路線」→ (制約詳細化) →「中華人民共和国の鉄道路線」→ (制約詳細化) →「中国の鉄道路線」のように、個別に見たときに、どちらの制約が詳細なのかがはっきりしないために、両方向ともに詳細化とつけてしまったものが存在した。後者については、2019 年 2 月 8 日現在の Wikipedia カテゴリ上では、「中華人民共和国の鉄道路線」→「中国の鉄道路線」の関係が削除されている。

4.2 クラス階層の分析

トップのカテゴリから、子カテゴリを持たないカテゴリまでの最短経路を深さとすると、まず、カテゴリ階層全体を見た

際の深さは、最大で 12、平均は 4.7 となっている (表 1)。「都市」→「各国の都市」→「ヨーロッパの都市」→「イタリアの都市」→ {「ナポリ」、「ローマ」、「ミラノ」、...} のような概念の分割が行われてから、最終的に topic カテゴリが 4,5 階層目で横並びになるため、平均はこのような値となっている。概念の種類 (set カテゴリと CS の set 部分) としては 17,026 種類あり、その中で、上記例の「都市」のように、制約の詳細化が多く起こったり、多種類の制約が付くことにより、複数回現れる概念 (10 件以上のもの) は 3,350 種類存在した。一番多いのは「人物」であり 45,514 件のカテゴリが存在し、Wikipedia カテゴリが、詳細化やトピックにより膨大になっており、クラス階層としては規模の大きくないことが伺える。次に、概念が同じものはひとまとまりとして数える。具体的には、包含関係が満たされない「クラス付加」「Instance of」「その他」の関係を辿らず（今回は簡易化するために「topic 包含関係」も辿らないとした)、かつ、「制約詳細化」の関係のときは深さを増やさないとして、深さを数える。上記例をとると、「都市」から「イタリアの都市」まで、全て概念「都市」の 1 階層となる。その結果、深さは最大で 9、平均で 3.5 となった。更にその中から、同じ深さにある同じ概念をまとめると深さは最大で 6、平均 2.5 となった (表 2)。「Instance of」等の関係を辿っていないため、topic カテゴリが抜けていることで全体のカテゴリ数は少なくなっている。また、上記のように、一部の概念は、「各国」→「ヨーロッパ」→「イタリア」のように、そこに付く制約の詳細化により分割され、深さが大きくなっているが、同じ概念のものをまとめ、最短経路をとると小さい階層としてまとめられることがわかる。深さが大きいものは、「リオデジャネイロオリンピック選手団」→「リオデジャネイロオリンピック日本選手団」のような CS と set の中間的なカテゴリを含む階層が含まれていた。現在これらはどちらも set となっていてクラス-サブクラスの関係に当たるが、「リオデジャネイロオリンピック (の) 日本 (の) 選手団」と考え「選手団」の階層としてひとまとまりに考えると、階層の深さは小さくなることが考えられる。

4.3 考察

分析の結果から分かるように、カテゴリの種類とその関係を整理することで、まず、殆どの循環構造が解消された。残ってしまったものの中で、「〜学」のような学問に関するカテゴリは、歴史や二国間関係等と同じようにサブクラスやインスタンスを広く取るように思われたので、改めて分類する必要があると考えた。その他循環構造が残る階層関係に関しては概念階層としては使わないとするのが良いと考えている。また、Wikipedia カテゴリ階層は様々なカテゴリが複雑に繋がりあっている構造に思えるが、概念に付けられた制約が詳細化されていることにより深さが増し、トピックにより異なる概念が繋がりあっているように見えていて、クラス-サブクラスに限定してクラス階層をみると、深さの小さい概念階層のような階層が存在していることが分かった。

5. まとめ

本研究では、整理されたデータを用いて Wikipedia カテゴリの再整理を行い、循環構造に関する検討ととクラス階層の分析から、これまでに提案してきた Wikipedia カテゴリに関する分類の枠組の有用性について議論した。分類について不十分な部分も見つかったが、概ね目的としていた階層を作ることができるデータができていることが本研究により明らかとなった。分類されたデータを複数人からのチェックを受け、洗練していくことで、カテゴリ階層も目的としていたものにより近づくことを期待している。

参考文献

[Bizer 09] Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S.: DBpedia - A crystallization point for the Web of Data, *Web Semantics: Science, Services and Agents on the World Wide Web*, Vol. 7, No. 3, pp. 154 – 165 (2009)

[Hoffart 13] Hoffart, J., Suchanek, F. M., Berberich, K., and Weikum, G.: YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia, *Artificial Intelligence*, Vol. 194, No. 0, pp. 28 – 61 (2013)

[Yoshioka 14] Yoshioka, M.: Analysis of Japanese Wikipedia Category for Constructing Wikipedia Ontology and Semantic Similarity Measure, in *Information Retrieval Technology 10th Asia Infomation Retrieval Societies Conference, AIRS 2014, Kuching, Malaysia, December 3-5, 2014 Proceedings*, pp. 470–481, Springer-Verlag GmbH (2014), LNCS8870

[玉川 10] 玉川 奨, 桜井 慎弥, 手島 拓也, 森田 武史, 和泉 憲明, 山口 高平?F 日本語 Wikipedia からの大規模オントロジー学習, 人工知能学会論文誌, Vol. 25, No. 5, pp. 623–636 (2010)

[中川 18] 中川 嵩教, 吉岡 真治?F 知識工学者のための日本語 Wikipedia のカテゴリ階層構造の再整理, 人工知能学会全国大会論文集, Vol. JSAI2018, pp. 2F402–2F402 (2018)

[中川 19] 中川 嵩教, 小板橋 佳晃, 吉岡 真治?FWikipedia カテゴリの構成要素に注目したカテゴリ階層の分析 (2019)

[藤原 12] 藤原 嵩大, 吉岡 真治?FWikipedia の階層関係を分析するためのカテゴリパターンの提案, 2012 年度人工知能学会全国大会 (第 26 回) 論文集 (2012), CD-ROM 2C1-NFC2-4

表 1: カテゴリ全体での深さ

深さ	カテゴリ数	例
0	9	人間、総記、社会
1	210	宇宙、工学、地名、人名
2	25,102	音楽、人工生命、亜鉛、経済学
3	13,449	林学、医師、亜鉛の化合物、バイオテクノロジー
4	45,724	アジアの環境、日本の食品、2017 年の地震、有毒植物
5	57,215	イタリアの博物学者、タイの河川、日本の海鮮料理
6	38,438	イタリア語のオペラ、協奏曲、日本の石橋、北海道の放送
7	18,812	北海道の警察署、明治大学、E-girls、王貞治
8	2,767	北海道のサッカーチーム、青森県の銀行、日本の男子プロゴルファー
9	232	イヌ科、北海道の社会人野球チーム、日本のプロ野球選手
10	55	コトドリ科、福島県ヒメオオ生息地、大森氏
11	40	オウチョウ科、ウグイス科、カラス科
12	1	ヨシキリ科

表 2: クラス階層の深さ

深さ	カテゴリ数	例
0	9	人間、総記、社会
1	351	法、教育、人物、企業
2	2,105	親族法、学校、医師、空港、サービス業
3	2,007	婚姻・離婚法、内科医、私立学校、スキー選手
4	439	婚姻法、リサーチ会社、電車、アイスダンス選手
5	49	リオデジャネイロオリンピック選手団、旧制高等学校、密閉式ドームスタジアム
6	11	リオデジャネイロオリンピック日本選手団、高等工業学校